OpinHuBank: szabadon hozzáférhető annotált korpusz magyar nyelvű véleményelemzéshez

Miháltz Márton

MTA Nyelvtudományi Intézet
mmihaltz@gmail.com

Az előadásban szeretnénk bemutatni az OpinHuBank véleményannotált korpuszt, melyet számítógépes érzelem-/véleményelemző rendszerek kutatásához, tanításához és teszteléséhez fejlesztettünk ki. A korpusz a META-SHARE[1] disztribúciós hálózaton keresztül szabadon hozzáférhető.

A számítógépes véleményelemzés (opinion mining) célja a szövegekben megjelenő szubjektív kifejezések – érzelmek, értékelések, álláspontok, vélemények, hiedelmek, gondolatok, érzések, ítéletek, spekulációk stb. – és azok polaritásának (pozitív vagy negatív), valamint célpontjának (melyik, a szövegben megnevezett entitásra irányul) feltárása [1]. A technológia a weben napi szinten elérhető milliós nagyságrendű szöveges forrás (blogok, fórumok, közösségi oldalak, hírportálok) felhasználásával olyan alkalmazási területeket tesz lehetővé, mint az üzleti döntések támogatása, brandek monitorozása, piaci elemzések, online közvélemény-kutatások stb.

Magyar nyelven Berend és Farkas a kettős állampolgárság témájában megnyilvánuló hozzászólók véleményének automatikus megállapítását tűzte ki célul gépi tanulásos megoldással [2]. Az OpinHu projekt (2009-2010) [3] a magyar mellett angol, német, kínai és arab nyelven működő, szabályalapú tartalomelemző rendszert fejlesztett ki, melyben érzelmi szótárként a Harvard General Inquirer lexikon lokalizációját (magyarul mintegy 4700 kifejezés), polaritást módosító mintákat (magyarul kb. 30 elem), valamint mély nyelvi elemzésre támaszkodó mintafelismerést alkalmaztak.

Az OpinHuBank projekt egy olyan kézzel annotált, magyar nyelvű erőforrás létrehozását tűzte ki célul, mely megfelel az alábbi célkitűzéseknek (hasonlóan [4]-hez):

- A korpusz mérete tegye lehetővé gépi tanuló rendszerek betanítását is (10,000 különböző annotált példa kontextussal együtt.)
- A korpusz nyelvezete a magyar híroldalak és blogok szövegét reprezentálja. A kész korpusz anyagának 27%-a blogokból, 73%-a hírportálok, hírügynökségek oldalairól származik
- A vélemények polaritása mellett a vélemények célpontjai is annotálva vannak, hogy a korpusz segítségével célpontokat detektáló módszereket is lehessen vizsgálni, fejleszteni. A célpontok a korpuszban névelemek (tulajdonnevek).

1 http://www.meta-share.eu

- Minden annotációs egységet több, egymástól független humán annotátor is lásson el jelöléssel (5 különböző annotátor).

A korpusz anyagának gyűjtéséhez az OpinHu projekt[2] adatbázisa szolgálta a kiindulópontot, mely több mint 500 meghatározó, rendszeresen frissülő hazai online forrásból (híroldalak, blogok, fórumok) több millió különböző szöveget tartalmaz a 2009-2012 közötti időszakból. Az annotációhoz előkészítést az alábbi lépésekben végeztük el.

Első lépésben a szövegeket automatikus mondathatár-felismerő, tokenizáló (hontoken), morfológiai elemző (hunmorph) és szófaji egyértelműsítő (hunpos), majd névelem-felismerő (huntag) eszközökkel dolgoztuk fel. A teljes adatbázis összes cikkének minden mondatából véletlenszerűen kiválasztottunk 12,000 különböző mondatot, amely tartalmazott legalább egy, PERSON (személynév) típusú entitást, valamint megfelelt néhány biztonsági kritériumnak (legalább 7 token hosszú, a végén található írásjel). Minden mondat minden különböző entitás-előfordulása egy-egy különböző annotációs egység lett (így, ha ugyanaz a név ugyanabban a mondatban többször is előfordul, akár mindegyik előfordulás különböző polaritásannotációt kaphat). Mivel az entitásokat elsősorban vélemények célpontjaként szerettük volna felhasználni, kiszűrtük azokat a példákat, ahol az entitás nagy valószínűséggel a mondatban megjelenő vélemény forrása volt. A leggyakoribb ilyen szerkezetek a mondaton belüli – egy szónál hosszabb – idézetek, az ilyet tartalmazó mondatokban az idézőjelen kívül (a főmondatban) előforduló neveket nem soroltuk az annotálandó egységek közé. Végül kézi ellenőrzéssel kiszűrtük azokat a példákat, amelyekben az entitásfelismerő hibázott (kategóriatévesztés), az így fennmaradó annotációs egységek közül kiválasztottuk az első 10,000 darabot.

Az annotációhoz a GeoX Kft. munkatársai létrehoztak egy webes felületet, ahol az annotátorok saját azonosítóikkal belépve, annotációs egységenként haladva végezhették el a munkát. Az 5 annotátor számára megfogalmazott irányelvben a következőket rögzítettük: a polaritások megítélésében csakis az entitás adott mondatban megfigyelhető státuszát vegyék figyelembe, ne a névhez kapcsolódó elsődleges szubjektív asszociációjukat, világról való tudásukat (ehhez segítség: képzeljék el, hogy a mondatban az entitás helyén az ő nevük szerepel, hogyan tetszene így a mondat?); ne a teljes mondat polaritását vegyék figyelembe, ez lehet különböző a benne szereplő kérdéses entitás polaritásától; ha egy mondatban egy névhez pozitív és negatív vélemény is kapcsolódik egyszerre, semleges polaritást jelöljenek; ha egy entitás polaritása nem dönthető el rövid gondolkodás után/bizonytalan, legyen semleges a polaritása.

Az elkészült, szabadon hozzáférhető korpusz CSV fájl formátumban tartalmazza az annotációs egységeket, melyek az alábbi elemekből állnak: az egység azonosítója, a mondat azonosítója, a mondatot tartalmazó szöveg eredeti URL-je, a mondat szövege, a célpont entitás szövege, a célpont entitás kezdőpozíciója és hossza a mondatban (tokenek), az 5 annotátor jelölései (-1: negatív, +1: pozitív, 0: semleges).

Az OpinHuBank munkálatai a CESAR projekt[3] támogatásával készültek el.

2 https://sites.google.com/a/geox.hu/opinhu/

3 http://cesar.nytud.hu/

Hivatkozások

1. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In: Proceedings of HLT/EMNLP 2005 (2005)
2. Berend, G., Farkas, R.: Opinion Mining in Hungarian based on textual and graphical clues. In: Proceedings of the 4th Intern. Symposium on Data Mining and Intelligent Information Processing. Santander (2008)
3. Miháltz M.: OpinHu: online szövegek többnyelvű véleményelemzése. In: VII. Magyar Számítógépes Nyelvészeti Konferencia. Szeged (2010) 14–23
4. Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., van der Goot, E., Halkia, M., Pouliquen, B., Belyaeva, J.: Sentiment Analysis in the News. In: Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'2010). Valletta, Malta (2010) 2216–2220